Stance Classification of Twitter Debates: The Encryption Debate as A Use Case

Aseel Addawood
Illinois Informatics Institute
University of Illinois at Urbana-Champaign
aaddaw2@illinois.edu

Jodi Schneider
School of Information Sciences
University of Illinois at Urbana-Champaign
jodi@illinois.edu

Masooda Bashir
School of Information Sciences
University of Illinois at Urbana-Champaign
mnb@illinois.edu

ABSTRACT
Social media have enabled a revolution in user-generated content. They allow users to connect, build community, produce and share content, and publish opinions. To better understand online users' attitudes and opinions, we use stance classification. Stance classification is a relatively new and challenging approach to deepen opinion mining by classifying a user's stance in a debate. Our stance classification use case is tweets that were related to the spring 2016 debate over the FBI's request that Apple decrypt a user's iPhone. In this "encryption debate," public opinion was polarized between advocates for individual privacy and advocates for national security. We propose a machine learning approach to classify stance in the debate, and a topic classification that uses lexical, syntactic, Twitter-specific, and argumentative features as a predictor for classifications. Models trained on these feature sets showed significant increases in accuracy relative to the unigram baseline.

CCS Concepts
• **Human-centered computing~Social networking sites** • **Computing methodologies~Natural language processing** • *Computing methodologies~Supervised learning*

Keywords
Stance Classification; Supervised Machine Learning; Natural Language Processing; Argumentative Features.

1. INTRODUCTION
Researchers have turned to user-generated content in social media as a source of information to explain many aspects of human experience [1]. Due to the often textual nature of online users' self-disclosure of their opinions and views, social media platforms present a unique opportunity to analyze shared content and, in particular, how controversial topics are argued. Continuous growth of online data has led to large amounts of information becoming available for others to explore and understand. For instance, Twitter has grown dramatically since its introduction over a decade ago to become one of the world's most popular social media platforms. Today, more than 288 million people actively use the site on a monthly basis [2].

#SMSociety'17, July 28-30, 2017, Toronto, ON, Canada

ACM ISBN 978-1-4503-4847-8/17/07…$15.00
http://dx.doi.org/10.1145/3097286.3097288

Automatic techniques, such as sentiment analysis and opinion mining, have allowed researchers and business people to determine the different viewpoints expressed in social media text (e.g., [3]). As their main task, these approaches assign a polarity score to an opinion that is presented in an online format. Although it is important to determine whether a user's opinion is positive or negative, it is even more essential to determine the user's position toward a specific topic [4].

Stance classification offers complementary information to sentiment analysis. Given a collection of debate-style discussions on a controversial topic, stance classification seeks to identify a user's attitudes toward the topic. This can support the identification of the user's affiliation with social or political groups, help develop better user-targeted recommendation systems, or tailor a user's information preferences to match his or her ideologies and beliefs [5-9]. Automatic stance classification can be used in applications such as information retrieval, text summarization, opinion summarization, and textual entailment.

Over the last decade, there has been active research in modeling stance. However, most of the work has focused on congressional debates [10] or debates in online forums [6,8,11,12]. Compared to these domains, Twitter is a much more challenging domain for stance prediction. Tweets are written in an informal format; they do not follow any guidelines or rules for the expression of opinions. Many messages contain unconventional syntax and spelling, which present a significant challenge to attempts at extracting meaning [13,14]. In this work, we investigate whether two argumentative features are beneficial for ideological stance classification and detect stance in one ideological debate—encryption in the United States, as discussed on Twitter following a high-profile event.

This particular online debate was kindled after the San Bernardino, California terrorist attack, which occurred in December 2015 [15]. For weeks following the attack, Apple Inc., one of the most well-known technology companies in the U.S., refused to create a "backdoor" that would give the Federal Bureau of Investigation (FBI) access to the encrypted iPhone of the alleged terrorist. Apple's refusal to comply with the FBI request gave rise to what we call the "encryption debate." This debate found its way into the mainstream media and became a popular topic of social media debate for months.

It provoked reactions from IT experts, politicians, and technologists as well as the public. Although this debate continues both offline and online, in this study we focus on the online encryption debate that occurred on Twitter from January 1st through March 31st, 2016. We selected this date range since

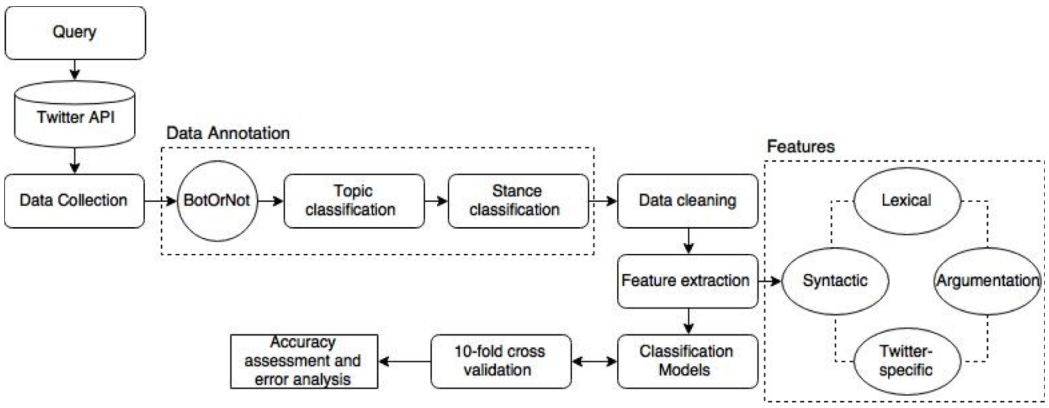


Figure 1. Project workflow.

it included tweets from the debate before and after a federal judge ordered that Apple unlock the iPhone on February 16, 2016 [15]. We were motivated to choose this use case because the tension between individual right to privacy and national security has long been of interest to philosophical, political, and technological debates. Those who favor national security argue that good citizens who have "nothing to hide" should not fear government surveillance and that law enforcement should have access to their information whenever necessary. Those who favor individual right to privacy argue for limiting government surveillance and access to personal information. As our mobile devices contain increasingly sensitive information and intricate details about our lives, the debate over whether information from these devices should be made available to law enforcement has become heated. Thanks to technological advances, many mechanisms have been developed to secure information to prevent unauthorized access. One of the most robust mechanisms, cryptography (i.e., encryption), allows messages to be sent confidentially. It is the use of the Advanced Encryption Standard that makes the iPhone such a formidable device to crack.

In this paper, we explore whether classifying stance in an ideological debate can determine how frequently each position is expressed in Twitter and what attitudes users express. We also explore which features can enhance the stance classification task. We describe a novel benchmark dataset of tweets that we labeled by both the topic of discussion and the user's stance towards that topic. The annotation is based on the stance that a user has expressed toward one of two topics: individual right to privacy and national security. Compared to our earlier work on argumentation mining of tweets [16], we use an additional layer of manual annotation to indicate the stance expressed about the main topics of discussion (described in Section 3.3 below) and perform a detailed analysis on the annotation results from both human annotators and automatic prediction. As we discuss in Section 3, we found that the argumentativeness of the tweet and its tone are suitable features for predicting the stance of the tweet. Figure 1 summarizes the workflow of this study.

Section 2 discusses related work in stance classification; Section 3 describes the data and corpus analysis; Section 4 discusses the experimental setup and feature selection; Section 5 presents the experimental results; Section 6 discusses the findings; and Section 7 concludes the paper and proposes future directions.

2. RELATED WORK

Supervised machine learning has been used in almost all of the current approaches to stance classification. One of the first studies related to stance classification dealt with perspective identification. Lin, Wilson, and Hauptmann [17] used articles from the Bitter-Lemons website, which discusses the Palestinian-Israeli conflict from each side's point of view, to train a system to perform automatic perspective detection on sentence and document levels. Later, Anand et al. [6] deployed a rule-based classifier with several features such as unigrams, bigrams, punctuation marks, syntactic dependencies, and the dialogic structure of posts from a competitive debating site. Their results ranged from 54% to 69% accuracy.

Somasundaran and Wiebe [18] created a lexicon for detecting argument trigger expressions and subsequently leveraged it to identify arguments. These extracted arguments, together with sentiment expressions and their targets, were used in a supervised learner as features for stance classification. This experimental work included both argument and sentiment features from four datasets—abortion, creationism, gun rights, and gay rights—each containing news articles from a wide variety of sources. Their overall accuracy result was 63.93%. Murakami and Raymond [19] identified general user opinions in online debates, distinguishing between global positions (opinions on a topic) and local positions (opinions on previous remarks). By calculating the degree of disagreement between any two users from the link structure and the text of each pair of their adjacent replies. Faulkner [20] investigated the problem of detecting document-level stance in student essays; their key features are (1) stance-taking clauses (in a generalized format that tracks long-distance dependencies, which they call part-of-speech-generalized stance proposition subtrees); and (2) reuse of words from the essay prompt. Sobhani, Inkpen, and Matwin [21] detected and classified stance starting by extracting online news comments using topic modeling.

To date, stance classification research has mainly focused on specific domains and mediums. Only a few studies have explored stance classification on social media. For example, Rajadesingan and Liu's study [22] used Twitter-based stance classification. The authors proposed a retweet-based label propagation method which starts from a set of known opinionated users and labels the tweets posted by the people in their retweet network. By contrast, in this work, we focus on detecting stance from a single tweet starting from a set of labeled tweets. Mohammad, Kiritchenko, Sobhani, Zhu, and

Cherry's [23] study also used Twitter as a dataset for stance classification. Their aim was to determine user stance (favor, against, or no position) in tweets on five selected topics: abortion, atheism, climate change, feminism, and Hillary Clinton. This dataset was made available for SemEval 2016, with two tasks. Task A was a traditional supervised classification task where 70% of the annotated data for a target is used as training and the rest for testing. The highest classification F-score for Task A was 67.82, with 19 teams participating. For Task B, test data was all of the instances for a new target (not used in Task A) and no training data was provided. The highest F-score for Task B was 56.28 with 9 teams participating. The dataset was offered to task participants without any context such as conversational structure or tweet metadata, which made classification challenging. In contrast, our approach for determining stance in this study takes into consideration tweet metadata (e.g., number of followers) as well as tweet labels that indicate a specific topic identified for the encryption debate.

3. DATA

3.1 Data Acquisition

In this research, we use publicly available social media data from Twitter. The initial dataset was originally gathered to investigate the classification of argumentative tweets [16]. This dataset was composed of 3,000 tweets from the encryption debate which we collected and then hand-annotated as we describe below. First we collected every public post on Twitter from January 1, 2016 through March 31, 2016 sent from accounts that set English as their language: 531,633 tweets in total, which we collected using Crimson Hexagon [24], a social media analytics platform that provides paid firehose access. We then filtered this data in several ways. We manually removed 40 tweets that were in another language even though the accounts language was set to English. This left 531,593 tweets in our dataset. Since we were only interested in real human opinions (not social bots or Sybil accounts), we excluded any user with a 50% or greater probability of being a bot based on the Truthy BotOrNot algorithm [25]. Overall, 946 tweets by bot accounts were removed. The total number of tweets after all adjustments was 530,647 tweets.

3.2 Data Annotation

3.2.1 Codebook Development and Annotation Schema

We used a data-driven and theoretically grounded approach to develop a practical solution to stance classification. We randomly selected a small sample of our corpus, 30 tweets, for close reading done by the first author. Our annotation outline consisted of two segments: topic classification and stance classification for each tweet. These two tasks were performed manually in conjunction with each other. We used an iterative process for developing the codebook. Initially, we developed three stance classifications and three topic classes (from the three most frequently discussed topics relevant to the debate) which are information privacy, national security and right to encryption. Two human annotators were trained through discussions with the first author to label 100 tweets in each of three iterations which created a total of 300 tweets as the development set. After each iteration, we had an extensive discussion of the challenges and limitations of the codebook. The resulting analysis led to a final revision of the coding scheme and modification of the associated codebook.

Table 1 contains a short overview of the codebook, showing specific definitions and example tweets. For topic classification, the final codebook had two main topics: information privacy and national security. We excluded the topic 'right to encryption' from our codebook since we realized, after discussions with annotators, that it was too generic and could cover both information privacy and national security. Moreover, users' attitudes toward the encryption debate seemed skewed towards those who valued either individual privacy or national security more highly. We added two additional categories to incorporate other types of tweets: those that shared news without expressing opinions about the two main topics ('other'); and those that contained jokes or nonsense ('irrelevant'). The final category scheme thus had four topic classifications: 'individual privacy', 'national security', 'other', and 'irrelevant'. For stance classification, three possible positions toward each topic were considered: 'favor,' 'against,' and 'neutral.'

To start the annotation process of the 3,000 tweets, we instructed our two annotators to first annotate each tweet based on the topic to which it was most related (topic classification), and to then annotate the posting user's overall position toward the topic (stance classification). By the end of the three iterations, approximately 33% (990) of the 3,000 tweets was labeled by both coders. We used Cohen's Kappa [26] to measure inter-annotator agreement. Our annotation consisted of two separate tasks, and the inter-coder reliability was 81.30% for topic classification and 87% on stance classification. The unweighted Cohen's Kappa score was 70% for topic classification and 64% for stance classification.

3.2.2 Annotation Challenges

We faced many challenges while annotating the tweets. We categorized tweets placed into the 'other' category as neutral since they did not provide an opinion, opposing or favoring, any of the topics analyzed. In this classification, we did not consider the stance of the article that was linked. For example, the following tweet does not represent a stance or share an opinion about the debate; it only shares the title of a news article and a link to it:

Amazon backtracks, decides to bring encryption back to Fire OS https://t.co/gK0I4tXn9l #tech

We cannot be certain how users feel about items they choose to retweet, and we generally classified most retweets as neutral. For instance, a user's retweeting of a CNN news story about an exchange between Apple and the FBI was marked as neutral. However, when a user retweeted something that very clearly expressed a stance, we counted the tweet as having that stance. For example, someone retweeting Edward Snowden speaking out against encryption backdoors would be marked as having a stance in 'favor' of the topic 'individual privacy'. We classified tweets that were completely unrelated to the debate as irrelevant; we did not consider it necessary to evaluate a user's stance in an irrelevant tweet. Tweets categorized as irrelevant were later excluded from the dataset, because they had no impact on the classification.

Table 1. Excerpt from codebook

	Class	Description	Example
Topic	**National security**	Government should protect the state and its citizens against all kinds of "national" crises related to the public's/the whole nation's interests.	*"I'm against backdoors, not against trying to hack a murderous terrorists encrypted phone #encryption"*
	Individual privacy	"The right to be let alone" [40].	*"I'm oddly paranoid of people reading my phone over my shoulder. Some day I will need to design personally language for encryption."*
	Other	Tweets that don't talk about national security or individual privacy, but are somewhat related to encryption. OR Tweets that are copies of news article titles without any comments.	*"End to end encryption: when will it be universal as a safe communication mode?" "Tim Cook Wants a Government Commission to Settle the War Over iPhone Encryption https://t.co/NshUf43f9b #TechNews"*
	Irrelevant	Tweets that are completely unrelated to encryption: jokes, nonsense.	*"Apple: 'Okay, here's the deal. We'll give you backdoor encryption, but you have to go through iTunes.'"*
Stance	**Favor**	Tweets that support one of topics by reacting positively or showing positive sentiments toward the topic or expressing their agreement.	*"I'm oddly paranoid of people reading my phone over my shoulder. Some day I will need to design personally language for encryption."*
	Against	Tweets that oppose one of topics by reacting negatively or showing negative sentiments toward the topic or expressing their disagreement.	*"@QuadPiece But why encryption in the first place? It's not realistically more secure, it's just slower."*
	Neutral	Tweets that ask questions OR Tweets that neither support nor oppose any of the topics or that do not show any positive or negative sentiments toward the topic.	*"I'm really torn on this phone encryption issue. #justsaying"*

3.3 Corpus Analysis

We manually labeled 3,000 tweets in total. The distribution of topics over the three stance labels is illustrated in Table 2. Additionally, this table shows the number of occurrences of each topic in the corpus. We can see that 'individual privacy' had a higher number of tweets than 'national security'. Table 3 provides an overview of the stance labels in this corpus. From this table, we can see that the neutral stance classification has the highest value. This echoes previous research that found that users do not frequently use Twitter as a debating platform [27]. Rather, most individuals use Twitter as a venue to spread information and share links to web pages instead of using it as a platform through which to have conversations about controversial issues. The results also illustrate that very few tweets were classified as being 'against' one of the topics.

4. EXPERIMENTAL SETUP

4.1 Preprocessing

Due to character limits, Twitter users tend to use colloquialisms, slang, and abbreviations. They also often make spelling and grammar errors. Before discussing feature selection, we will briefly discuss how we compensated for these issues in data preprocessing. First, we tokenized tweets using the ARK Tweet NLP tokenizer [28]. This Twitter-specific tokenizer segments tweet features such as emoticons, hashtags, and mentions. We replaced emoticons with their sentiment polarity. Next, we replaced abbreviations with their whole word or phrase counterparts (e.g., 2night => tonight). We then removed duplicated vowels in the middle of words (e.g., haaaapy). Any letter occurring more than two times in a row was replaced with exactly two occurrences. Inspired by [16], this modification significantly reduced feature space. Finally, we lowercased all letters (e.g., ENCRYPTION => encryption) and removed URLs and mentions to other users, after first recording these features.

4.2 Features

Based on prior work [29,30], we chose four types of features: lexical, syntactic, Twitter-specific, and argumentation. Table 4 provides a summary of the features we extracted for each tweet. Below, we describe and explain the motivation for these feature sets. In the following sections, we propose a set of features to characterize stance in tweets. Much of our work uses lexical features, which can help find words that are both highly salient and highly informative in a text or text set. This process also entails the removal of a) non-content-bearing words that dominate with respect to the cumulative power-law distribution of word frequencies, and b) highly rare words in a collection. After preprocessing the data, we considered salient unigrams and bigrams, removing stopwords and removing any word with fewer than five occurrences. Previous work suggests that the unigram baseline can be difficult to beat for certain types of debates [18]. Thus, we used both unigrams and bigrams as

Table 2. Distribution of topics labels in the corpus

Topic classification	Class distribution	Percentage
Individual privacy	329	10.96%
National security	25	0.83%
Other	2,505	83.5%
Irrelevant	141	4.7%

Table 3. Distribution of stance labels in the corpus

Stance classification	Class distribution	Percentage
Favor	345	11.5%
Against	8	0.27%
Neutral	2,647	88.2%

Table 4. Feature types used in our model

Type	Feature	Description
Lexical	Unigram	Word count for each single word that appears in the tweet
	Bigram	Word count for every two words that appear in the tweet
Syntactic	Sentiment	Positive, negative, or neutral sentiment
	Subjectivity	Strong, weak, or neutral subjectivity
	Grammatical	Number of occurrences of noun, verb, adjective, preposition, adverb, and pronoun
Twitter-specific	Retweet	1.0 if the tweet is a retweet
	Title	1.0 if the tweet contains the title to an article
	Mention	1.0 if the tweet contains a mention to another user "@"
	Verified account	1.0 if the author has a "verified" account
	URL	1.0 if the tweet contains a link to a URL
	Followers	Number of people this user is following at posting time
	Following	Number of people following this user at posting time
	Posts	Total number of user's posts
	Hashtag	1.0 if the tweet contains a hashtag "#"
Argumentation	Argumentativeness	1.0 if the tweet is argumentative
	Source type	Type of source used in the tweet

features. We kept the top 500 unigrams and the top 300 bigrams according to the TF-IDF metric as shown in Equation 3.

$$TF(t) = tf(t,d) \quad (1)$$

$$IDF(t) = \log\left(\frac{|D|}{1+|\{d:t \in d\}|}\right) \quad (2)$$

$$tf - IDF(t) = TF * IDF \quad (3)$$

In these formulas, t is a term, d is the document in which t occurs, and D is the document space (collection of documents). Equation 1 shows the term frequency of word t, Equation 2 the inverse document frequency, and Equation 3 the TF-IDF score calculation for term t.

4.2.1 Syntactic Features

Syntactic features describe the re lationship between words and their roles in a sentence, as the subjectively connoted adjectives and other modifiers, sentiment, and the ratio of different parts of speech in a sentence. In natural language processing, these characteristics are standard features for machine learning.

Sentiment: After experimenting with other sentiment analysis dictionaries such as the Subjectivity Lexicon [31], we selected the sentiment labels provided by Crimson Hexagon [24], since it seemed to provide more accurate results than other sentiment analysis dictionaries.

Subjectivity: We used the MPQA Subjectivity Lexicon [31] to identify the subjectivity or objectivity of tweets.

Grammatical features: We used the NLTK part-of-speech tagger [32] to assign a single best-fitting part of speech (POS) to every token. We calculated POS diversity by finding the number of occurrences of each POS tag.

4.2.2 Twitter-Specific Stylistic Features

Twitter-specific features refer to characteristics unique to the Twitter platform that are associated with user accounts and the tweets sent from them, such as the number of followers, number of people followed, and the number of tweets the user has posted in the past. Twitter-specific features also include the presence or lack of URLs, mentions of other users, hashtags, and official account verification. These features were acquired using the Twitter API, and we treated them as part of the structure of the tweets, and thus necessary, for our analysis. Therefore, before preprocessing the data, we first calculated the number of occurrences of each of these features in a tweet and added them to the set of attributes.

4.2.3 Argumentation Features

We used the dataset provided by [16], which is labeled with argumentation and source type. We used these two labels as part of our feature set.

Argumentativeness: We used a simple argument model that an argument is comprised of only two components: a claim and associated supporting evidence. If the tweet presented an argument or shared an opinion about the debate, it was marked as argumentative, and otherwise, as not argumentative.

Source type: Source type refers to the type of evidence a user has given to support a particular position in a given debate. Six types of evidence were identified: 'news media accounts'; 'blog post'; 'picture'; 'expert opinion'; 'other types of evidence', and 'no evidence,' which referred to not having presented any evidence.

4.3 Imbalanced Class Distributions

It was not possible to control the class distribution by controlling the Twitter query, because determining the topic class and the stance class had to be manually determined as described in Section 3.3. But the imbalances shown in Tables 2 and 3 above could bias the classifier, i.e., the classes with fewer instances could be predicted incorrectly and with lower accuracy than classes with more instances. Previous studies have proposed various balancing strategies, including oversampling, undersampling, cost-sensitive learning, and a combination of these methods [33,34]. Previous work has shown that the combination of oversampling and undersampling techniques performs better than plain undersampling [35,36] and has a better outcome than cost-sensitive learning [37]. Therefore, to resolve imbalanced class distributions, we used a combination of two techniques: oversampling for classes with a small number of instances, and undersampling for classes with a large number of instances. For oversampling, we used the Synthetic Minority Oversampling Technique (SMOTE) [35]. SMOTE is one of the most accepted approaches for solving the problem of imbalanced data, and has better performance compared to oversampling with replacement [35]. Its main function is to create new minority class examples by interpolating several

minority class instances that occur together. In this method, new instances are synthetically created using k-nearest neighborhoods. Based on the number of cases in each class, a range from 500% to 900% was chosen using k=5 to minimize the risk of overfitting the classifier. After that we used random undersampling to reduce the size of large classes with a ratio of 5:1. Finally, we randomized the data to reduce the likelihood of overfitting the training data. Table 5 shows the new class distributions after balancing the dataset. The size differences between classes have been minimized.

5. EXPERIMENTAL RESULTS

The primary aim of our study was to determine the stance of tweets towards a certain topic. We used a multi-classification task to classify each tweet as having a stance in 'favor,' 'against,' or 'neutral'. As a first step, we compared classifiers that have frequently been used in related work: Naïve Bayes (NB) as used in [11]; Support Vector Machines (SVM) as used in [38]; and Decision Trees (DT, J48) as used in [30]. For all approaches, we used WEKA data mining software [39]. Before training, all features were ranked by their information gain [40]. Information gain is presented in Equation 4.

$$InfoGain(Class, Attribute) = P(Class) - P(Class|Attribute) \quad (4)$$

Features with information gain of less than 0 were excluded. All results were subjected to 10-fold cross validation. For assessing prediction accuracy, we used the standard metrics of precision, recall, and F-measure. The results for each feature set and classifier are listed in Tables 6 and 7.

5.1 Classification

Our first goal was to classify the topics related to encryption i.e. national security and individual privacy. Table 6 shows a summary of the classification results. The best results were achieved by using DT, which resulted in an F1 score of 93.7%. Our second goal was to classify tweets based on their stance toward a predefined topic. Adding a bigram feature to the baseline did not increase performance; however, adding argumentation features to the combination increased the performance by 10% for SVM. Table 7 shows a summary of our classification results. To achieve them, we created a baseline model by using the top salient unigrams. A baseline needed to be established so that we could assess the influence of added features on the models. The best overall performance was achieved by using SVM, which resulted in a 90.4% F1 score with lexical and argumentation-mining features added to the baseline. Moreover, combining all the features slightly decreased SVM and NB performance but did not change DT results substantively.

Table 5. Number of instances for each class after balancing

	Class	Class distribution after balancing
Topic	Individual privacy	329
	National security	150
	Other	750
Stance	Favor	345
	Against	80
	Neutral	480

5.2 Feature Analysis

To identify and rank the most informative attributes of each feature, we calculated information gain (Eq. 4). The top 10 features with the largest weight (magnitude) with respect to each class are listed in Table 8. As shown in the table, the most informative feature in all classes was argumentativeness. Among Twitter-specific features, retweets appeared in both 'favor' and 'neutral' stances, as well as in the 'individual privacy', and 'other' topics. Among syntactic features, Crimson Hexagon sentiment features were informative for the 'favor' stance, as well as for the 'individual privacy', and 'other' topics. Moreover, among syntactic features we found that the most informative grammatical features for the topic of national security were preposition, adjective, verb, and adverb. Figure 2 shows the top 10 lexical features for each class. Among these features, we found that all classes had a combination of unigram and bigram features. The unigram 'I' was one of the top informative features in all classes except the topic of national security. From Table 8, we can see that features as 'privacymatters', 'spying on', and words related to standing up for encryption are negatively associated with the 'against' stance. Moreover, sentiment bearing words, i.e., 'should' are a good indicator of 'neutral' stance where it is negatively associated with the 'favor' stance. For topic classification, we can see from Table 8 that the word 'encryption' is negatively correlated with the topic of individual privacy. Based on these findings, we conclude that using both lexical and argumentation features was beneficial for this task. As our analysis of the top informative attributes shows, the structure of sentences, grammatical indices, subjective words, and argumentativeness of tweets were useful for predicting the stance and topic.

Table 6. Topic classification results of three classifiers using 10-fold cross validation

Feature set	DT			SVM			NB		
	Pre.	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
Unigram (Baseline)	90.3	90.3	90.2	88.3	88.4	88.3	84.6	83.4	83.8
All features	93.7	93.7	**93.7**	93.2	93.2	**93.2**	85.9	84.5	**84.9**

Table 7. Stance classification results of three classifiers using 10-fold cross validation

Feature set		DT			SVM			NB		
		Pre.	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
Lexical	Unigram (Baseline)	76.3	76.1	76.2	81	81	81	79.1	78.7	78.8
	Unigram + Bigram	76.5	76.6	76.5	81.7	81.7	81.6	78.8	78.2	78.4
Lexical+ Syntactic		75.7	75.7	75.7	82.9	82.9	82.8	81.4	81.0	81.1
Lexical + Argumentation		77.8	77.8	**77.8**	90.4	90.4	**90.4**	83.4	82.8	**82.9**
All features		77.6	77.6	77.6	83.8	83.8	83.8	79.4	79.3	79.2

Table 8. Most and least informative features (*non-lexical features in italic*)

	Class	Most informative Features	Least informative Features
Topic	Individual privacy	*argumentativeness*, *retweet*, I, *sentiment*, I'm, stand, support, I stand, harder, I support	encryption from, encryption fight, encryption engineers, encryption security, encryption for, encryption debate, encryption I, encryption so, encryption technology, encryption support
	National security	*preposition*, *adjective*, *verb*, *adverb*, harder, than, committee, them, *Argumentativeness*, in ISIS	requiring encryption, really hope, powerful encryption, protect us, protest against, phone mass, privacy apples, one don't, one consider, outta luck
	Other	*argumentativeness*, *retweet*, *sentiment*, *verb*, *preposition*, *adjective*, harder, I, *adverb*, I'm	internet commerce, internet like, i trust, iphone might, iphone encryption, layer encryption, law enforcement, key encryption, keys secure, keep getting
Stance	Favor	*argumentativeness*, *retweet*, *sentiment*, encryption that, I, create an, not have, sides, I believe, unconstitutional	sense privacy, so called, shocked that, should get, should too, sick of, side encryption, side of, side w, cloud storage
	Against	not have, *noun*, I believe, see both, believe apple, unconstitutional, sides in, both sides, is unconstitutional, create an	standup privacymatters, spying on, stand behind, stance how, stance in, standing up, stand by, stand up, stand with, cloud storage
	Neutral	*argumentativeness*, *retweet*, I, verb, but I, Apple should, not have, case but, should, is unconstitutional	retweet to, secure because, right don't, right to, secure at, right on, san Bernardino, rock solid, same encryption, cloud storage

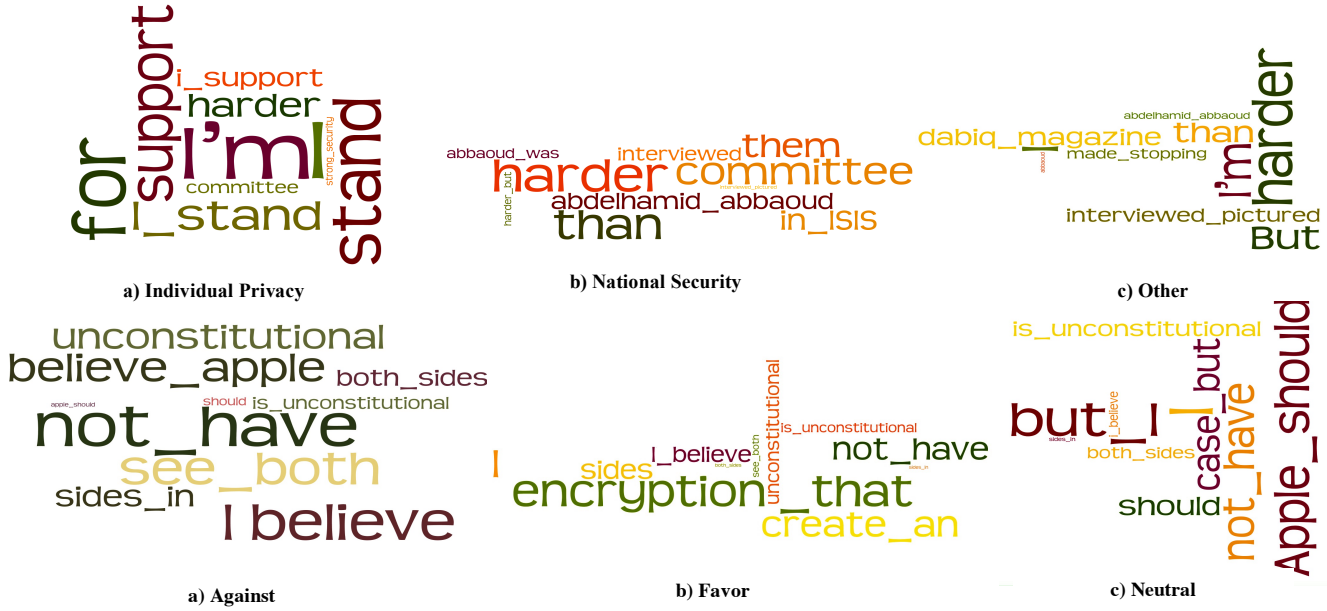
6. Error Analysis

In addition to analyzing contributions within and among features' classes, we also studied each classifier's confusion matrix to find patterns in misclassifications. For stance classification, we chose the SVM's confusion matrix because of its comparatively higher accuracy with lexical and argumentation feature sets. Table 10 shows the number of classified instances per stance class, rendered in percentages. As Table 10 shows, 'favor' and 'neutral' classes were the most misclassified classes. This result was consistent with our human annotators' feedback. They found it difficult to distinguish between tweets that favored a topic and tweets that did not take a stance toward the debate.

In particular, it was challenging to distinguish between those who had a clear opinion about the topic versus those who were just making a joke about it. For topic classification, we chose the Decision Tree's confusion matrix. Table 9 shows the classified instances per topic class, rendered in percentages. As the table shows, the 'individual privacy' and 'other' topic classes were the most misclassified. To further analyze these prediction errors, we randomly selected 30 tweets from different classes, removed the labels, and asked the same two human annotators to label them again. Their unweighted Cohen's Kappa scores were 81.37% for stance classification and 70.4% for topic classification. This finding shows that some tweets were hard to categorize and suggests that understanding the intended meaning of the tweets might be needed to solve this problem. Based on our discussion with the human annotators, we believe that being able to see the whole conversation preceding the tweet and being familiar with the content of the shared URLs could lower these errors.

7. DISCUSSION

In this study, we developed a theoretically grounded and data-driven classification schema, related codebook, corpus annotation, and prediction model for detecting stance in tweets from the "encryption debate." Our data annotation and analysis



a) Individual Privacy b) National Security c) Other

a) Against b) Favor c) Neutral

Figure 2. Word cloud of the Most informative lexical features for each class.

Topic classes (top row) and stance classes (bottom row).

Table 9. Topic classification confusion matrix of DT classifier (by percentage)

	Individual privacy (%)	National security (%)	Other (%)
Individual privacy	**87.5**	0.60	11.85
National security	1.33	**94.6**	4
Other	3.2	0.53	**96.26**

Table 10. Stance classification confusion matrix of SVM classifier (by percentage)

	Against (%)	Favor (%)	Neutral (%)
Against	**95**	2.5	2.5
Favor	0	**85.8**	14.2
Neutral	0	7.1	**92.91**

procedure showed that most individuals use Twitter as a venue for spreading information and links to webpages rather than as a platform through which to take clear positions about controversial issues. In Table 2, the distribution of topic label results show that 'individual privacy' had a higher number of tweets compared to 'national security.' We think this bias toward 'individual privacy' may have happened because people are more confident tweeting about their personal right to privacy rather than the more public responsibility to maintain national security. It may also indicate that people are more willing to share their opinions if they thought that their audience agreed with them. These results can be compared to a recent Pew research study [41] about Edward Snowden's 2013 revelations of widespread government surveillance of Americans' phone and email records. The survey showed that 86% of Americans were willing to have an in-person conversation about the surveillance program, but just 42% of Facebook and Twitter users were willing to post about it on those platforms. The distribution of stance label results in Table 3 confirm previous research which showed that users do not frequently use Twitter as a debating platform [27]. The results also illustrate that very few tweets were classified as being 'against' one of the topics. This may indicate that Twitter users do not take 'against' stances as frequently as stances in 'favor' of controversial topics, especially if those topics are morally and not scientifically based.

To build classifiers, we worked with four sets of features: lexical, syntactic, Twitter-specific, and argumentation. We trained three commonly used types of classifiers: Support Vector Machine, Decision Tree, and Naïve Bayes. We built a baseline model using top unigrams, gradually added other feature types, and measured the incremental contribution of each type. For topic classification, we only compared the baseline to the combination of all features because of the need to limit our research scope for this paper. The classification results (Table 6) showed that the combination of all four sets of features was most beneficial for the DT classifier, with which the results improved from 90.2% for the baseline to 93.7%. The Naïve Bayes scores for F1, recall, and precision were lower than those for the other two classifiers. Our results indicate a 20% improvement in F-measure score compared with previous research [9, 18, 22]. We believe that the unique combination of features used in the classification as Twitter-specific features, sentiment, and argumentation facilitated these improvements. The stance classification results (Table 7) show that the SVM classifier outperformed the other two training algorithms and achieved the best overall performance. It did so by using a combination of lexical and argumentation features, which led to a performance that improved from the 81% baseline F1 score to the 90.4% final model F1 score.

The comparison of the top attributes of each class revealed that one argumentation feature, which indicated whether or not a tweet is argumentative, is the best indicator for stance classification. This may indicate that when a tweet is argumentative it denotes that the user expresses a stance toward the topic. Moreover, the retweeting behaviour was observed in tweets that have an in 'favor' or 'neutral' stance only. Also, the same retweeting behaviour was observed in tweets discussing 'individual privacy' and 'other' topics. This may indicate that users on Twitter are more comfortable sharing information in 'favor' or 'neutral', towards 'individual privacy', and 'other' topics, but not toward 'national security.' One limitation of our study is that we understand that our dataset may not be representative of the overall opinions of Twitter users online. As our sample shows, only 1% of the annotated dataset was about national security while few tweets had an 'against' stance. However, we believe that these results still provide some information about Twitter users' attitudes towards the encryption debate. We found that some lexical features are very indicative of the topic class. In the case of the individual privacy topic, for example, the top lexical features were 'for', 'stand', 'support', and 'I stand'. These features indicate a very strong position toward the topic, in contrast to the national security topic where the first personal pronoun did not appear as one of the top features. This result may indicate that users are less comfortable expressing their own opinions when the topic involves national security, or that they are more comfortable discussing a personal matter, such as their privacy, rather than a collective issue as the national security of the whole country. We also conducted an error analysis of misclassified instances, finding that tweets related to 'favor' and 'neutral' stances, as well as 'individual privacy' and 'other' topics, were the most challenging to classify. This occurred because of the challenge of classifying short texts that do not follow any guidelines or rules for the expression of opinions and the challenge of distinguishing sarcasm from earnest opinions.

8. CONCLUSION AND FUTURE WORK

The analysis of social media content has been studied extensively. There are many challenges to opinion-mining social media content, because online users' expressions are written informally, and so may include sarcasm, spelling mistakes, unconventional grammar, and slang words and expressions [13, 14]. Several works have begun to develop tools and computational models for tweet-level opinion and sentiment analysis. Although opinion mining and sentiment analysis can identify whether a user expresses a positive or negative emotion regarding a topic, these techniques may not capture a user's stance (in favor or against a given position) on the topic. Stance classification has been introduced to address this gap. Although as yet under-investigated, stance classification has seen growing interest in recent years, as this technique can be advantageous, particularly in support of decision-making. In order to detect online users' attitudes and stances on a given issue, we used Twitter data related to the recent Apple and FBI encryption debate. In this paper, we presented the task of automatically classifying stance on social media for users discussing controversial topics like the recent FBI and Apple encryption debate utilizing unique feature sets. We classified two predefined topics related to the debate and built a dataset of 3,000 manually annotated tweets related to these topics. Our

subsequent analysis, motivated by the research presented in [16], found that SVM classifiers trained with lexical and argumentative features were best at capturing stances taken toward different topics expressed on social media. While previous work has considered classifying stance without any tweet context, we show that using various features such as the sentiment and the argumentativeness of the tweet support the identification of the stance of the tweet and can lead to significant improvements in stance classification.

As stated previously, working with social media data has some challenges and limitations. Annotating tweets related to a controversial topic such as the encryption debate requires annotators who not only understand the English language used and its informing cultures, but who also understand the encryption debate as a whole. Another challenge of annotating the data was related to the language and structure of tweets, in which users tend to use informal and incoherent text. In addition, it is important to note that although our classification achieved a high score in our selected debate topic, these results may not be generalizable to other domains without further investigation. Understanding public opinions and attitudes towards controversial topics may help scholars, law enforcement officials, and policy-makers develop better policies and guidelines. People's attitudes and behaviours related to privacy are highly contextualized in the digital age. While many scholars have conceptualized information privacy in various disciplines, investigations of individual users' attitudes and behaviours towards information privacy and national security remain limited. The dataset developed in this paper will be used in future research to develop a better understanding of users' attitudes towards the encryption debate: ultimately that may help enhance current privacy policies and guidelines. Given the growing significance of the role social media is playing in our world, studying stance classification can be beneficial for instance, in identifying electoral issues and understanding how public stance is shaped [23]. One implication of our research is that it suggests that it is possible to understand who frequently participates in controversial discussions on social media. Moreover, correlating users' stance with their sentiments and demographics may help further describe users' behaviour online. Also, predicting a user's stance toward a given issue can support the identification of social or political groups, help develop better recommendation systems, or tailor users' information preferences to their ideologies and beliefs. Additionally, it may provide engineers and designers with new ways of improving the design and users' acceptability of current privacy-enhancing technologies.

In future work, we hope to improve our results with more intelligent features for representing context, discourse, rhetorical structure, and dialogic structure, such as capturing irony and sarcasm. Another area to explore in future work is analyzing tweets based on the whole conversation, instead of just a single tweet, to get a better understanding of users' different opinions. Another line of research to pursue in the future is to develop a system that can detect the different stances users have regarding a controversial topic, i.e., explore how people decide what the sides (two, three, more) are in a given debate. A controversial topic may generate many different and nuanced stances, even on the same general side of a debate.

9. REFERENCES

[1] Anatoliy Gruzd and Melissa Goertzen. 2013. Wired academia: Why social science scholars are using social media. In *Proceedings of the 2013 46th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 3332-3341.

[2] Wei Wang, Ivan Hernandez, Daniel A Newman, Jibo He and Jiang Bian.2016.Twitter analysis: Studying US weekly trends in work stress and emotion. *Applied Psychology*, 65, 2 (2016), 355-378.

[3] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. ACL, 79-86.

[4] Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu and Colin Cherry. 2016. Semeval-2016 Task 6: Detecting stance in tweets. *Proceedings of SemEval*, 16 (2016).

[5] Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. ACL, 399-409.

[6] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. ACL, 1-9.

[7] Jean Mark Gawron, Dipak Gupta, Kellen Stephens, Ming-Hsiang Tsou, Brian Spitzberg and Li An. 2012. Using group membership markers for group identification in web logs. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI, 467-470.

[8] Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*. ACL, 1348-1356.

[9] Minghui Qiu, Liu Yang and Jing Jiang. 2013. Modeling interaction features for debate side clustering. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. ACM, 873-878.

[10] Matt Thomas, Bo Pang and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. ACL, 327-335.

[11] Simone Teufel and Marc Moens.2002.Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28, 4 (2002), 409-445.

[12] Marilyn A Walker, Pranav Anand, Robert Abbott and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 592-596.

[13] Antonio Reyes, Paolo Rosso and Davide Buscaldi. 2012.From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74 (2012), 1-12.

[14] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, 704-714.

[15] Dave Lee. 2016, February 18. Apple vs the FBI—A plain English guide. Retrieved September 16, 2016 from http://www.bbc.com/news/technology-35601035.

[16] Aseel A. Addawood and Masooda N. Bashir. 2016. “What is your evidence?” A study of controversial topics on social media. In *Proceedings of the 3rd Workshop on Argument Mining*. ACL, 1-11.

[17] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe and Alexander Hauptmann. 2006. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*. ACL, 109-116.

[18] Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. ACL, 116-124.

[19] Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics*. ACL, 869-875.

[20] Adam Faulkner.2014.Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. In *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference*. AAAI, 174-179.

[21] Parinaz Sobhani, Diana Inkpen and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*. ACL, 67-77.

[22] Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in Twitter debates. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 153-160.

[23] Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51, 4 (2015), 480-499.

[24] S. Etlinger and W. Amand. 2012. Crimson Hexagon [Program documentation]. Retrieved September 15, 2016 from http://www.crimsonhexagon.com/wp-cotent/uploads/2012/02/CrimsonHexagon_Altimeter_Webinar_111611.pdf.

[25] Clayton A. Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini and Filippo Menczer. 2016. BotOrNot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 273-274.

[26] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 1 (1960), 37-46.

[27] Laura M. Smith, Linhong Zhu, Kristina Lerman and Zornitsa Kozareva. 2013. The role of social media in the discussion of controversial topics. In *Proceedings of the 2013 International Conference on Social Computing (SocialCom)*. IEEE, 236-243.

[28] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL2011)*. ACL, 42-47.

[29] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 183-194.

[30] Carlos Castillo, Marcelo Mendoza and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 675-684.

[31] Theresa Wilson, Janyce Wiebe and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. ACL, 347-354.

[32] Steven Bird, Ewan Klein and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc., Sebatstopol, CA.

[33] Nitesh V. Chawla. 2005. Data mining for imbalanced datasets: An overview in *Data mining and knowledge discovery handbook*. Springer, 853-867.

[34] Sotiris Kotsiantis, Dimitris Kanellopoulos and Panayiotis Pintelas. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30, 1 (2006), 25-36.

[35] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16 (2002), 321-357.

[36] Rezvaneh Rezapour and Jana Diesner. 2017. Classification and detection of micro-level impact of issue-focused documentary films based on reviews. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1419-1431.

[37] Nitesh V. Chawla, Nathalie Japkowicz and Aleksander Kotcz. 2004. Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6, 1 (2004), 1-6.

[38] Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28, 7 (2012), 991-1000.

[39] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H Witten. 2009.The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11, 1 (2009), 10-18.

[40] Danny Roobaert, Grigoris Karakoulas and Nitesh V. Chawla. 2006. Information gain, correlation and Support Vector Machines in *Feature extraction*. Springer. 463-470.

[41] Keith Hampton, Lee Rainie, Weixu Lu, Maria Dwyer, Inyoung Shin and Kristen Purcell. 2014. Social Media and the ‘Spiral of Silence’. Retrieved September 10, 2016 from http://www.pewinternet.org/2014/08/26/social-media-and-the-spiral-of-silence/.